SN Computer Science (2020) 1:124
https://doi.org/10.1007/s42979-020-00135-w

ORIGINAL RESEARCH

Visual Linguistic Model and Its Applications in Image Captioning

Ravin Kumar[1]

Received: 27 March 2020 / Accepted: 30 March 2020

Abstract
Image captioning is a well-known task of generating textual description of a given image. Research work on this problem statement requires efforts in both computer vision and natural language processing domains to obtain better quality image descriptions. In this paper, we are proposing a new deep learning approach to generate image captions. In this approach, we generate a sequence of visual embeddings for objects and their relationships present in the image. These visual embeddings are arranged in a particular manner and are then supplied to the encoder part of an attention-based sequence-to-sequence model. In the final step, we receive the generated image captions from the decoder part of our sequence-to-sequence model. We tested its performance on MSCOCO Dataset, and the obtained results suggested that our model generates better image captions for MSCOCO testing dataset.

Keywords Image captioning · Image description · MSCOCO Dataset · Scene understanding · Deep learning

Introduction

Deep learning has accelerated the speed of research work in various fields. Development of convolutional neural network-based deep learning architectures [1–3] has played a major role in providing better solutions to computer vision-based problems [4–6]. Similarly, varieties of recurrent neural networks [7–9] are used to solve problems in natural language processing [10–12]. Although there are some problem statements which exist in the intersection of computer vision and natural language processing domain, one such problem statement is image captioning where the objective is to generate textual descriptions for the input image. Designing better deep learning-based solutions for this problem statement relies heavily on developments done in both domains which makes this problem statement even more tough to deal with.

In this paper, we are proposing a new deep learning-based architecture named "visual linguistic model" to solve image captioning problem efficiently and obtain better textual descriptions of images.

Related Works

Lin et al. [13] presented MSCOCO Dataset for benchmarking the performance on image captioning task. Xu et al. [14] proposed an end-to-end deep learning-based solution containing convolutional neural networks for feature extraction and recurrent neural network with attention generating image captions. You et al. [15] combined both top-down and bottom-up approach of image captioning and proposed a semantic attention model. Aneja et al. [16] suggested that better accuracy in image captioning can be obtained by using convolutional neural networks instead of using recurrent neural network for generating textual description. Anderson et al. [17] combined both top-down and bottom-up attentions to obtain better performance in visual question answering and image captioning tasks.

This article is part of the topical collection "Advances in Computational Approaches for Artificial Intelligence and Image Processing" guest edited by Bhanu Prakash K N and M. Shivakumar.

✉ Ravin Kumar
ravin.kumar.cs.2013@miet.ac.in

[1] Department of Computer Science, Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh 250005, India

Published online: 16 April 2020

Visual Linguistic Model

Our proposed deep learning approach takes advantage of the language translation abilities of sequence-to-sequence-based architectures. In our approach, a new mechanism is used for generating a sequence of embeddings that can represent the entire scene along with relationship of objects present in the scene. These sequences are then fed to the encoder part of an attention-based sequence-to-sequence model. In the decoder part, image captions are provided during the training phase, to help the deep learning model learn to generate better textual descriptions of the input image. Our approach converted the image captioning problem into language translation problem, and this is the reason why the proposed approach is named as "visual linguistic model."

Generating Embedding Sequences for Sequence-to-Sequence encoder

Firstly, an object detection model is used for detecting objects present in the input image, and then after receiving bounding boxes of detected objects, new images of objects are created from the input image. Finally, each image is resized to a fixed dimension of H × W. YOLO [8] is used for object detection because of its ability to provide a better end-to-end deep learning solution for detecting objects. Then, an autoencoder is trained using the input image and newly formed object images (Fig. 1).

After training, encoder part of the autoencoder generates embedding for the original image and its newly formed images of objects (Fig. 2). Some other pretrained autoencoders or other approaches [18] can also be used to generate embeddings, as long as they provide consistently better embedding of the image.

After gathering all the embeddings of an original input image, and objects present in it, some specific rules are followed to generate an effective input sequence for the encoder part of sequence-to-sequence model. These rules create a parent–child relationship tree from the object coordinates generated by the object detection model (YOLO in our case) for our original input images.

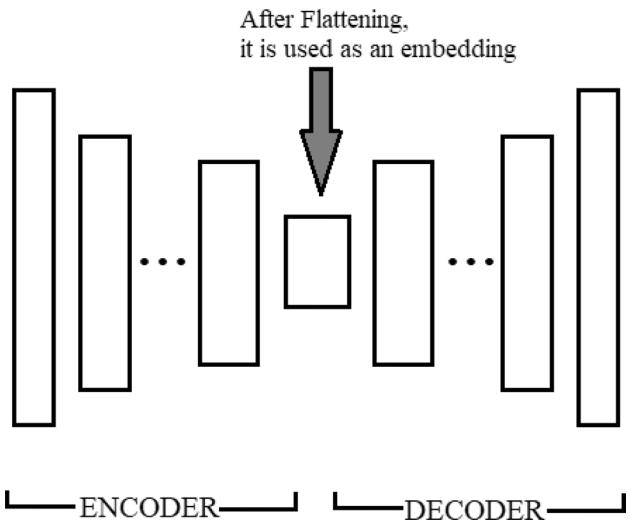

Fig. 2 Abstract representation of an autoencoder

We define a percentage threshold TH for intersection area. When objects bounding boxes have intersections with each other above TH value, the object with larger intersection area become a child for the other intersecting object (Fig. 3). One can also calculate the image depth of objects to define a better parent–child relationship. Original input image will always be the root node of this relationship tree. Instead of only relying on object bounding boxes, one can also utilize segmentation approach [6] to improve the calculation of the intersection area.

The generated tree is read in the bottom-up fashion to create the embedding sequence. While reading the tree, embeddings of the read objects are kept on including the embedding sequence. The last entry in the embedding sequence always contains the embedding of the original input image because this arrangement of reversed input sequences of encoder provides better performance. Although in case, one wants to experiment with non-reversed input sequence, reversing this input embedding sequence will provide the effect of reading tree in a top-down fashion which can then be fed to the encoder.

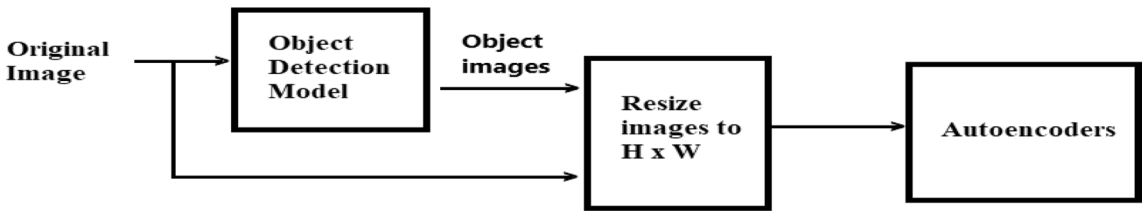

Fig. 1 Resized original image along with its object images send to the autoencoder

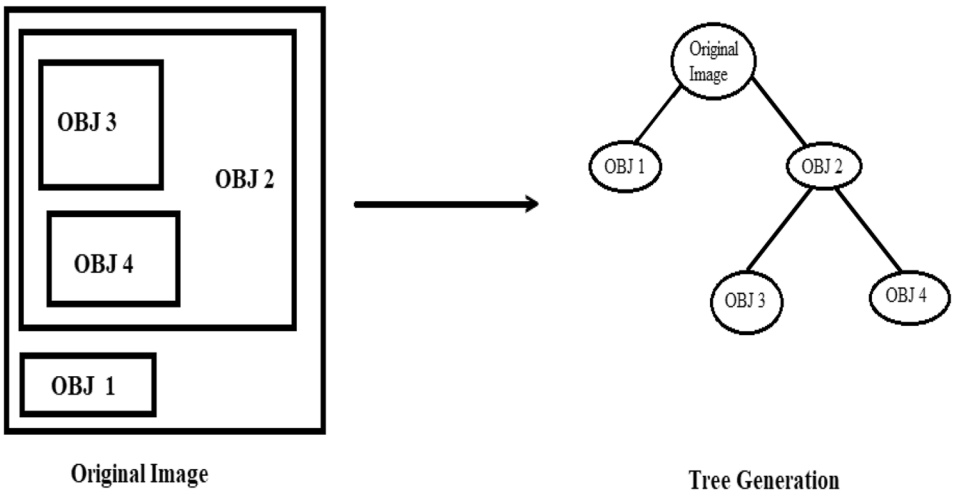

Fig. 3 Sample representation of tree generation

Overall Architecture of Visual Linguistic Model

An attention-based sequence-to-sequence model is trained using our tree generated embedding sequence in the encoder part with its associated textual description as input to the decoder part. To avoid any confusion, workflow diagram of our entire approach is provided to help better understand the visual linguistic model (Fig. 4). Source code for modules of deep learning architectures used in visual linguistic approach is available in our Github repository [19].

The main advantage of using visual linguistic model is that it converts the image captioning problem into a problem of language translation (Fig. 5). It allows the visual linguistic model to take advantage of already available end-to-end deep learning solutions of language translation to solve the image captioning problem (Fig. 6).

Dataset Used

There are many content-rich datasets available for image captioning problem, but for utilizing advantages of visual linguistic model, one requires training data for object

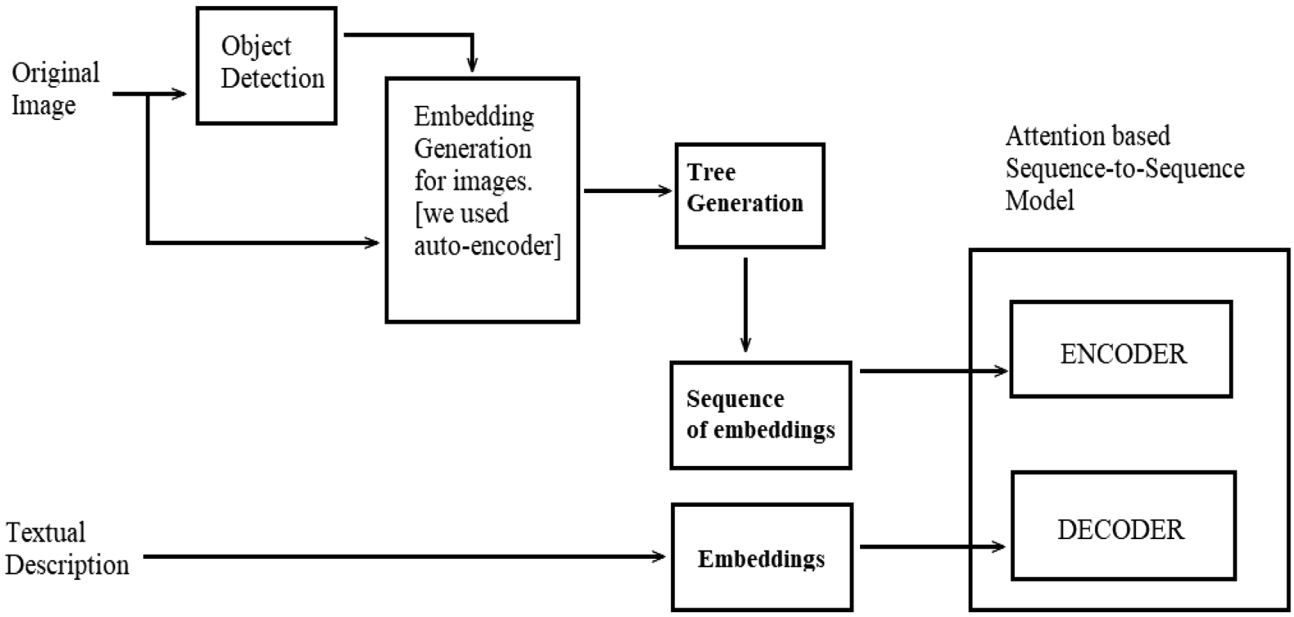

Fig. 4 Overall architecture of visual linguistic model

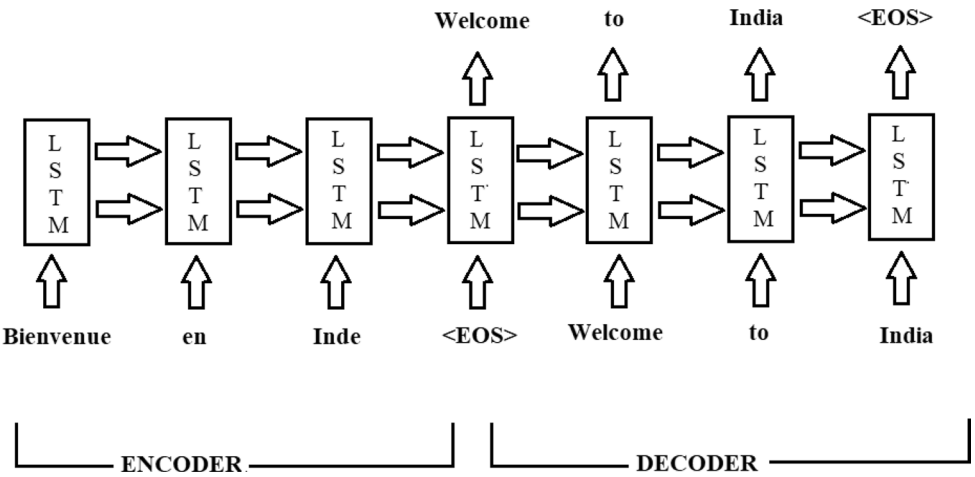

Fig. 5 Sequence-to-sequence model for language translation

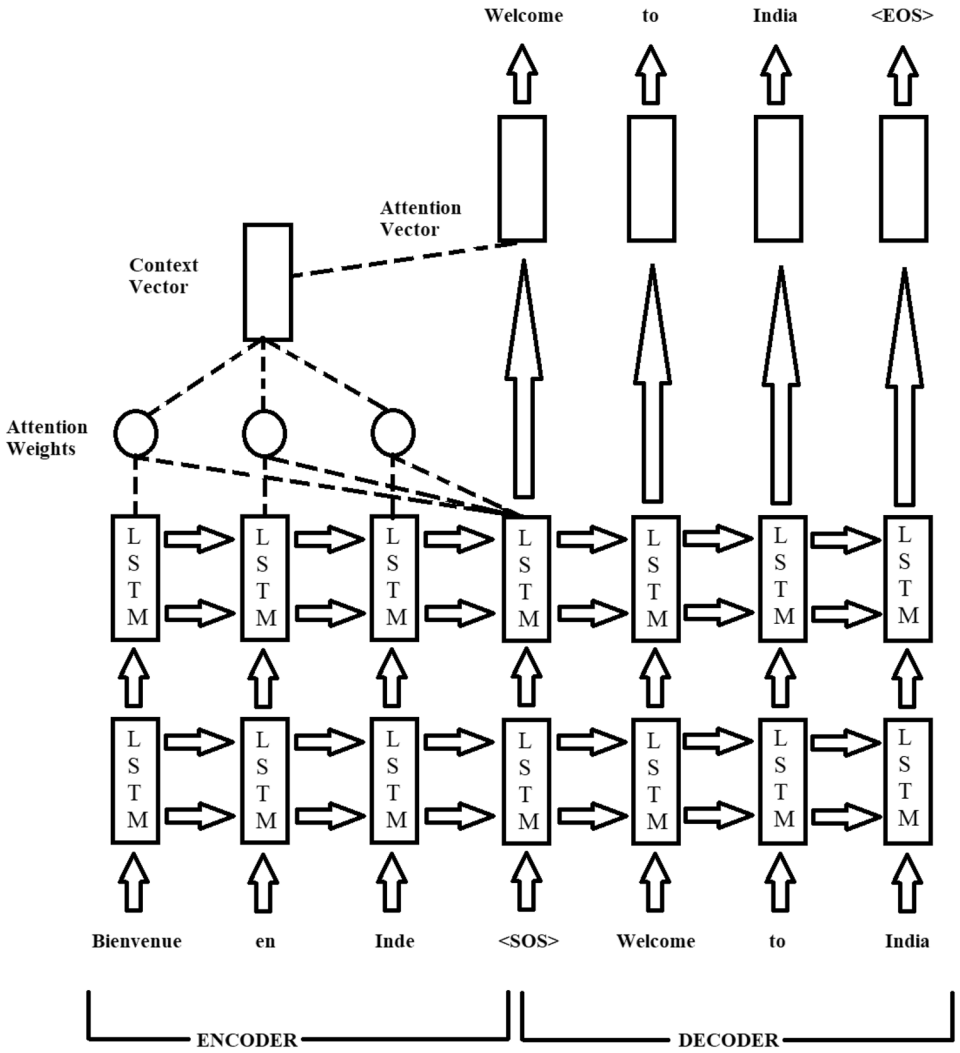

Fig. 6 Attention-based sequence-to-sequence model for language translation

detection, and it is a plus if segmentation data are also available because it helps in better tree generation which later provides the embedding sequence for the encoder part of sequence-to-sequence model.

MSCOCO Dataset contained various challenges on image captioning, segmentation, and detection (Table 1) and because of this, it became an ideal dataset for performing the comparative analysis with our visual linguistic approach.

Table 1 MSCOCO Dataset details

S. no.	Task	Challenge year
1	Image captioning	2015
2	Object detection	2015, 2016, 2017, 2018
3	Keypoints detection	2016, 2017, 2018
4	Stuff segmentation	2017, 2018
5	Panoptic segmentation	2018

We used MSCOCO Dataset [13] for testing the accuracy of our approach. It can be easily seen that without much hyper-parameter tuning, visual linguistic model easily outperformed the top solutions submitted on MSCOCO image captioning challenge [20] on CIDEr-D [21], ROUGE-L [22], METEOR [23], and BLEU [24] scores. However, we believe much better accuracy can be obtained with the explicit tuning of training hyper-parameters.

Comparative Analysis

We performed comparative analysis on the top performing models present on the MSCOCO image captioning leaderboard. Instead of using BLEU-1, BLEU-2, or BLEU-3 scores, we used only BLEU-4 score in both c40 and c5 cases.

The performance over c40 (Table 2) and c5 (Table 3) is separately analyzed in a tabular format. In both c40 and c5 comparative analyses, visual linguistic model has performed better than the top performing models and scored better in BLEU-4, METEOR, ROUGE-L, and CIDEr-D performance matrices. It is believed that with the better tuning of hyper-parameters, our model can provide much better performance.

Table 2 Comparative analysis for c40

S. no.	User/model	BLEU-4	METEOR	ROUGE-L	CIDEr-D
1	Visual linguistic	0.721	0.392	0.784	1.310
2	MIL-HDU	0.718 (1)	0.383 (1)	0.746 (1)	1.300 (1)
3	Lun	0.708 (5)	0.381 (4)	0.741 (4)	1.286 (2)
4	TencentAI.v2	0.701 (8)	0.377 (8)	0.737 (7)	1.278 (3)
5	h-p-hl	0.709 (4)	0.382 (3)	0.741 (3)	1.272 (4)
6	SRCB-ML_Lab	0.713 (3)	0.373 (14)	0.731 (19)	1.267 (5)
7	Dajiangyou3	0.697 (10)	0.372 (17)	0.736 (8)	1.265 (6)
8	pp2	0.694 (13)	0.375 (13)	0.735 (9)	1.257 (7)
9	AnonymousModel	0.703 (6)	0.379 (6)	0.738 (6)	1.256 (8)
10	Dajiangyou2	0.687 (20)	0.370 (22)	0.731 (17)	1.255 (9)
11	Cap_ann3	0.695 (12)	0.378 (7)	0.733 (11)	1.252 (10)
12	AnonymousTeam	0.692 (14)	0.372 (16)	0.731 (16)	1.251 (11)
13	CapJK	0.698 (9)	0.377 (9)	0.733 (12)	1.247 (12)
14	TingYao	0.691 (15)	0.373 (15)	0.729 (20)	1.246 (13)
15	ETA-Transformer	0.702 (7)	0.380 (5)	0.739 (5)	1.244 (14)
16	AnonymousResearcher	0.715 (2)	0.382 (2)	0.744 (2)	1.243 (15)
17	Anony_ultra	0.676 (26)	0.370 (18)	0.725 (23)	1.241 (16)
18	Ttry_speak	0.696 (11)	0.376 (10)	0.732 (15)	1.240 (17)
19	LiuDaqing	0.690 (17)	0.370 (19)	0.731 (18)	1.238 (18)
20	Iva_cococaption	0.690 (16)	0.375 (11)	0.734 (10)	1.236 (19)
21	Cascaded-Agents	0.681 (23)	0.369 (26)	0.726 (22)	1.234 (20)
22	wzn0828	0.668 (36)	0.367 (32)	0.720 (32)	1.224 (21)
23	BrianJ	0.688 (19)	0.367 (30)	0.720 (31)	1.223 (22)
24	fkxssaa	0.674 (27)	0.370 (21)	0.722 (27)	1.220 (23)
25	ak_txt	0.672 (29)	0.369 (24)	0.723 (26)	1.217 (24)
26	Schen_umn_vips	0.690 (18)	0.370 (20)	0.733 (13)	1.210 (25)
27	AdamTong	0.687 (21)	0.375 (12)	0.732 (14)	1.209 (26)
28	Panderson_msr	0.685 (22)	0.367 (31)	0.724 (25)	1.205 (27)
29	Discriminative	0.666 (37)	0.366 (38)	0.719 (34)	1.204 (28)
30	Image_caption_a	0.670 (33)	0.363 (45)	0.719 (35)	1.201 (29)

Table 3 Comparative analysis for c5

S. no.	User/model	BLEU-4	METEOR	ROUGE-L	CIDEr-D
1	Visual linguistic	0.401	0.297	0.598	1.291
2	MIL-HDU	0.399 (1)	0.290 (1)	0.593 (1)	1.283 (1)
3	Lun	0.392 (4)	0.288 (2)	0.588 (3)	1.261 (2)
4	TencentAI.v2	0.386 (7)	0.286 (5)	0.587 (4)	1.254 (3)
5	h-p-hl	0.390 (5)	0.287 (4)	0.586 (5)	1.250 (5)
6	SRCB-ML_Lab	0.397 (2)	0.284 (9)	0.585 (8)	1.253 (4)
7	Dajiangyou3	0.385 (8)	0.282 (14)	0.586 (7)	1.238 (7)
8	pp2	0.384 (10)	0.284 (8)	0.584 (9)	1.240 (6)
9	AnonymousModel	0.385 (9)	0.286 (7)	0.583 (10)	1.233 (8)
10	Dajiangyou2	0.378 (16)	0.281 (17)	0.582 (11)	1.227 (12)
11	Cap_ann3	0.373 (21)	0.281 (18)	0.574 (26)	1.212 (17)
12	AnonymousTeam	0.380 (14)	0.282 (15)	0.582 (14)	1.229 (11)
13	CapJK	0.374 (19)	0.281 (20)	0.574 (25)	1.211 (18)
14	TingYao	0.382 (11)	0.283 (12)	0.582 (12)	1.232 (9)
15	ETA-Transformer	0.389 (6)	0.286 (6)	0.586 (6)	1.221 (13)
16	AnonymousResearcher	0.396 (3)	0.287 (3)	0.590 (2)	1.231 (10)
17	Anony_ultra	0.373 (23)	0.281 (16)	0.578 (18)	1.218 (15)
18	Ttry_speak	0.373 (20)	0.280 (24)	0.573 (28)	1.206 (21)
19	LiuDaqing	0.379 (15)	0.281 (19)	0.582 (15)	1.216 (16)
20	Iva_cococaption	0.380 (13)	0.284 (10)	0.582 (13)	1.219 (14)
21	Cascaded-Agents	0.373 (22)	0.280 (23)	0.577 (20)	1.211 (19)
22	wzn0828	0.368 (27)	0.279 (26)	0.574 (27)	1.203 (23)
23	BrianJ	0.376 (17)	0.279 (27)	0.575 (24)	1.209 (20)
24	fkxssaa	0.368 (31)	0.282 (13)	0.577 (22)	1.205 (22)
25	ak_txt	0.367 (32)	0.280 (21)	0.576 (23)	1.195 (24)
26	Schen_umn_vips	0.381 (12)	0.279 (25)	0.581 (16)	1.187 (25)
27	AdamTong	0.375 (18)	0.283 (11)	0.580 (17)	1.183 (26)
28	Panderson_msr	0.369 (26)	0.276 (32)	0.571 (33)	1.179 (27)
29	Discriminative	0.363 (34)	0.277 (30)	0.571 (35)	1.179 (28)
30	Image_caption_a	0.368 (29)	0.275 (33)	0.572 (32)	1.179 (29)

Sample of Generated Captions

In this section, we have provided captions generated over sample images, along with the captions generated using traditionally used deep learning methods. During analysis of generated captions, it was found that our proposed approach generated better detailed descriptions of an image (Fig. 7).

Conclusion

Visual linguistic provides a new approach to create an embedding sequence from an image to solve image captioning. These sequences are used in the encoder part, and its equivalent textual descriptions are sent to the decoder part of an attention-based sequence-to-sequence model. Our deep learning model allows utilizing many deep learning-based solutions for the language translation problem to an

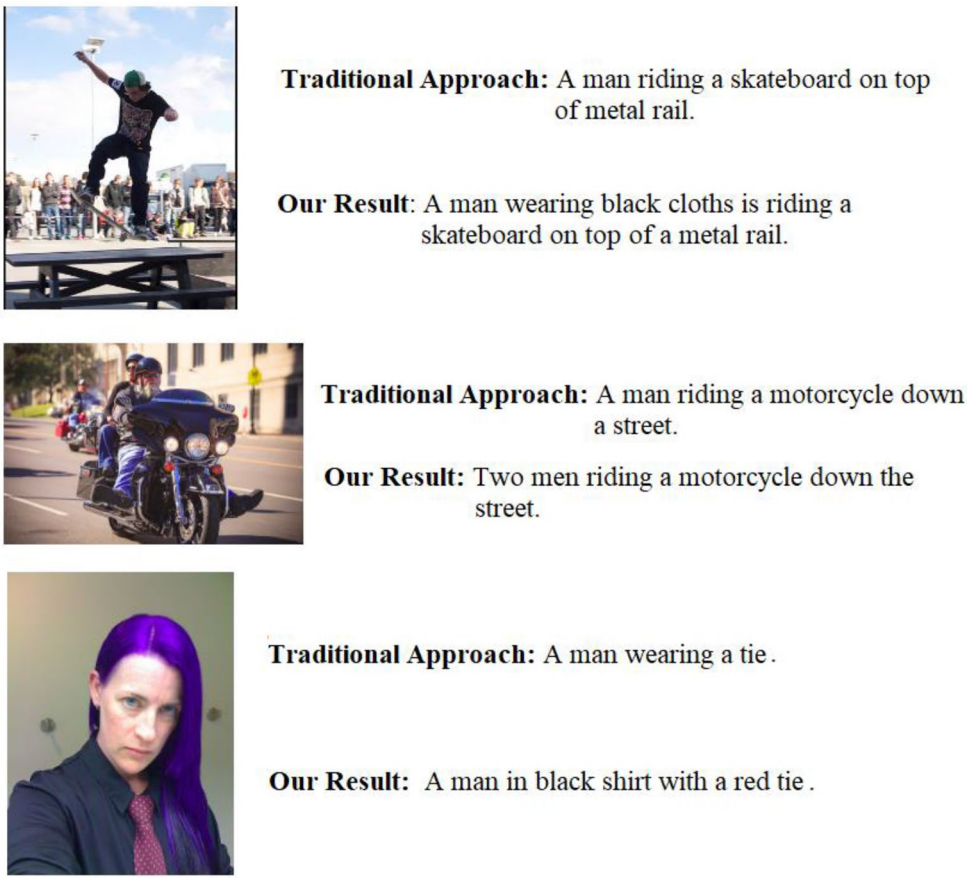

Fig. 7 Sample of generated image captions using visual linguistic model

image captioning problem. The presence of sequence-to-sequence architectures in visual linguistic approach can also be found useful for problems similar to image captioning including visual question answering and designing visual search engines.

Funding No Institutional funding is received to perform this research work.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv :1409.1556.
2. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.
3. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4700–8.
4. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 779–88.
5. Jackson AS, Bulat A, Argyriou V, Tzimiropoulos G. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 1031–9.
6. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3431–40.
7. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
8. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555. 2014 Dec 11.
9. Dey R, Salemt FM. Gate-variants of gated recurrent unit (GRU) neural networks. In: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). New York: IEEE; 2017. p. 1597–600.
10. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. 2014. p. 3104–12.
11. Nallapati R, Zhou B, Gulcehre C, Xiang B. Abstractive text summarization using sequence-to-sequence rnns and beyond. 2016. arXiv preprint arXiv:1602.06023.

12. Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. 2018. arXiv preprint arXiv:1812.09449.
13. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: common objects in context. In European conference on computer vision. Cham: Springer; 2014. p. 740–55.
14. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. 2015. p. 2048–57.
15. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2016. p. 4651–9.
16. Aneja J, Deshpande A, Schwing AG. Convolutional image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 5561–70.
17. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 6077–86.
18. Geng J, Fan J, Wang H, Ma X, Li B, Chen F. High-resolution SAR image classification via deep convolutional autoencoders. IEEE Geosci Remote Sens Lett. 2015;12(11):2351–5.
19. Github Repository. https://github.com/mr-ravin/visual-linguistic-model. Accessed 1 Mar 2019.
20. MSCOCO Captioning challenge. https://competitions.codalab.org/competitions/3221#results,last. Accessed 14 Mar 2019.
21. Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 4566–75.
22. Lin CY. Rouge: a package for automatic evaluation of summaries. Text summarization branches out. 2004.
23. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. p. 65–72.
24. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics; 2002. p. 311–8.